

# Nonadaptive Group Testing with Random Set of Defectives

Arya Mazumdar

**Abstract**—In a *group testing* scheme, a series of *tests* are designed to identify a small number  $t$  of defective items that are present among a large number  $N$  of items. Each test takes as input a group of items and produces a binary output indicating whether any defective item is present in the group. In a non-adaptive scheme the tests have to be designed in one-shot. In this setting, designing a testing scheme is equivalent to the construction of a *disjunct matrix*, an  $M \times N$  binary matrix where the union of supports of any  $t$  columns does not contain the support of any other column. In principle, one wants to have such a matrix with minimum possible number  $M$  of rows.

In this paper we consider the scenario where defective items are random and follow simple probability distributions. In particular we consider the cases where 1) each item can be defective independently with probability  $\frac{t}{N}$  and 2) each  $t$ -set of items can be defective with uniform probability. In both cases our aim is to design a testing matrix that successfully identifies the set of defectives with high probability. Both of these models have been studied in the literature before and it is known that  $\Theta(t \log N)$  tests are necessary as well as sufficient (via random coding) in both cases.

Our main focus is explicit deterministic construction of the test matrices amenable to above scenarios. One of the most popular ways of constructing test matrices relies on *constant-weight error-correcting codes* and their *minimum distance*. In particular, it is known that codes result in test matrices with  $O(t^2 \log N)$  rows that identify any  $t$  defectives. We go beyond the minimum distance analysis and connect the *average distance* of a constant weight code to the parameters of the resulting test matrix. Indeed, we show how distance, a pairwise property of the columns of the matrix, translates to a  $(t+1)$ -wise property of the columns. With our relaxed requirements, we show that using explicit constant-weight codes (e.g., based on algebraic geometry codes) we may achieve a number of tests equal to  $O(t \frac{\log^2 N}{\log t})$  for both the first and the second cases. While only away by a factor of  $\frac{\log N}{\log t}$  from the optimal number of tests, this is the best set of parameters one can obtain from a deterministic construction and our main contribution lies in relating the group testing properties to average and minimum distances of constant-weight codes.

**Index Terms**—Group testing, Disjunct matrices, Constant-weight codes, Deterministic construction

## I. INTRODUCTION

Combinatorial search is an old and well-studied problem. In the most general form it is assumed that there is a set of  $N$  elements among which at most  $t$  are *defective*. This set of defective items is called the *defective set* or *configuration*.

The author is with the College of Information and Computer Sciences, University of Massachusetts Amherst, MA 01002, email: [arya@cs.umass.edu](mailto:arya@cs.umass.edu). A part of this work was presented in the International Symposium on Algorithms and Computation, 2012 [30] and took place while the author was in Massachusetts Institute of Technology. This research is supported in part by NSF CCF-1318093, NSF CCF-1642658, and NSF CCF-1642550.

To find the defective set, one might test all the elements individually for defects, requiring  $N$  tests. Intuitively, that would be a waste of resource if  $t \ll N$ . On the other hand, to identify the defective configuration it is required to ask at least  $\log \sum_{i=0}^t \binom{N}{i} \approx t \log \frac{N}{t}$  yes-no questions. The main objective is to identify the defective configuration with a number of tests that is as close to this minimum as possible.

In the *group testing* problem, a *group* of elements are tested together and if this particular group contains any defective element the test result is positive. Based on the test results of this kind one *identifies* (with an efficient algorithm) the defective set with minimum possible number of tests. The schemes (grouping of elements) can be adaptive, where the design of one test may depend on the results of preceding tests. For a comprehensive survey of adaptive group testing schemes we refer the reader to [12].

In this paper we are interested in non-adaptive group testing schemes: here all the tests are designed together. If the number of designed tests is  $M$ , then a non-adaptive group testing scheme is equivalent to the design of a binary *test matrix* of size  $M \times N$  where the  $(i, j)$ th entry is 1 if the  $i$ th test includes the  $j$ th element; it is 0 otherwise. As the test results, we see the Boolean OR of the columns corresponding to the defective entries.

Extensive research has been performed to find out the minimum number of required tests  $M$  in terms of the number of elements  $N$  and the maximum number of defective elements  $t$ . The best known lower bound says that it is necessary to have  $M = \Omega(\frac{t^2}{\log t} \log N)$  tests [13], [16]. The existence of non-adaptive group testing schemes with  $M = O(t^2 \log N)$  is also known for quite some time [12], [22]. On the other hand, for the adaptive setting, schemes have been constructed with as small as  $O(t \log N)$  tests, optimal up to a constant factor [12], [21].

In the literature, many relaxed versions of the group testing problem have been studied as well. For example, in [17], [43] recovery of a *list* of items containing the true defectives is suggested (list-decoding superimposed codes). This notion was revisited in [8], [24] as list-disjunct matrix and in [19], where it was assumed that recovering a large fraction of defective elements is sufficient. There are also information-theoretic models for the group testing problem where the test results can be noisy [2] (also see [5], [7]). In other versions of the group testing problem, a test may carry more than one bit of information [4], [23], or the test results are threshold-based (see [9] and references therein). Algorithmic aspects of the recovery schemes have been studied in several papers. For example, papers [24] and [34] provide efficient recovery

algorithms for non-adaptive group testing.

Here as well, we consider two relaxed versions of the group testing problem – we want recovery to be successful with high probability assuming uniform distributions of the defective items. In the first scenario, each of the  $N$  items can be defective with probability  $\frac{t}{N}$ . This model of defectives, called **Model 1** throughout the rest of this paper, is as old as the group testing problem [11] and was rigorously defined in [38]. It is also the subject of very recent works such as [37]. We provide explicit construction of test matrices with  $O(t \log^2 N / \log t)$  tests for this situation. In the second scenario, we want the recovery to be successful for a very large fraction of all possible  $t$ -sets as defective configurations. This scenario, called **Model 2** throughout this paper, was considered under the name of *weakly separated design* in [29], [46] and [27]. It is known (see, [46]) that, with this relaxation it might be possible to reduce the number of tests to be proportional to  $t \log N$ . However this result is not constructive. Here also we provide explicit construction of test matrices with  $O(t \log^2 N / \log t)$  tests. Note that, this result is order-optimal when  $t$  is proportional to  $N^\delta$  for  $0 < \delta \leq 1$ .

In particular, our result leads to improvement over the construction of weakly-separated design from [18], whenever  $\log N \leq (\log t)^2$ . In [18], the total  $N$  items are partitioned and then a nonadaptive scheme for a smaller set of elements is repeated on each of the parts. It follows from a simple union bound that one would need  $O(t \log t \log N)$  tests for both the above random models to have high probability identification.

The repeated-block construction of [18] is analogous to repeating a good error-correcting code of small length to construct a long error-correcting code. Indeed, one can find the best linear error-correcting code of length  $\log n$  and then repeat that  $n / \log n$  times to construct a capacity-achieving code of length  $n$ . While this can be a first construction, it does not give any insight regarding the properties that are important for the problem. In an earlier conference version [30] of this paper, we showed that the properties of the distribution of Hamming distances of the columns of testing matrix can play a role in identification. While the result of [30] leads to suboptimal number of tests, we can use better concentration inequalities to arrive at improvements over it (see, Theorem 5). Our construction also turns out to give better parameters than the repetition scheme of [18], whenever  $\log N \leq (\log t)^2$ . Note that, this in particular include the regime where  $t$  varies as  $N^\delta$  for  $0 < \delta < 1$ , which is the premise of very recent works such as Scarlett and Cevher [37]. There is no apparent relation to the work of [18] with our techniques. In particular, our ideas cannot be viewed as an extension of repeated block construction.

We believe that our main contribution lies in 1) relating the group testing properties to the average Hamming distance between the columns of testing matrix and 2) using proper classes of explicit codes (such as Algebraic-Geometric codes) that satisfy the required properties of average and minimum distances.

Non-adaptive group testing has found applications in multiple different areas, such as, multi-user communication [3], [44], DNA screening [33], pattern finding [26] etc. It can be

observed that in many of these applications it would have been still useful to have a scheme that identifies almost all different defective configurations if not all possible defective configurations. The above relaxations form a parallel of similar works in compressive sensing (see, [6], [31]) where recovery of almost all sparse signals from a generic random model is considered.

A construction of group testing schemes from error-correcting code matrices and using code concatenation appeared in the seminal paper by Kautz and Singleton [25]. Code concatenation is a way to construct binary codes from codes over a larger alphabet [28]. In [25], the authors concatenate a  $q$ -ary ( $q > 2$ ) Reed-Solomon code with a unit-weight code to use the resulting codewords as the columns of the testing matrix. Recently in [35], an explicit construction of a scheme with  $M = O(t^2 \log N)$  tests is provided. The construction of [35] is based on the idea of [25]: instead of the Reed-Solomon code, they take a low-rate code that achieves the Gilbert-Varshamov bound of coding theory [28], [36]. Papers, such as [15], [45], also consider construction of non-adaptive group testing schemes.

In this paper we show that the explicit construction of [35] based on error-correcting codes works for both Model 1 and Model 2 and results in numbers of tests claimed above. Not only that, using explicit families of Algebraic-Geometric codes in conjunction with Kautz and Singleton construction we obtain test-matrices with the same performance guarantee.

#### A. Results and organization

The constructions of [25], [35] and many others are based on *constant-weight error-correcting codes*, a set of binary vectors of same Hamming weight (number of ones). The group-testing recovery property relies on the pairwise *minimum distance* between the vectors of the code [25]. In this work, we go beyond this minimum distance analysis and relate the group-testing parameters to the *average distance* of the constant-weight code. This allows us to connect the group testing matrices designed for random models of defectives to error-correcting codes in a general way (see, Thm. 2 and Thm. 3). Previously the connection between distances of the code and weakly separated designs was only known for the very specific family of *maximum distance separable* codes [27], where much more information than the average distance is evident.

Based on the newfound connection, for both Model 1 and Model 2, we construct explicit (constructible deterministically in polynomial time) families of non-adaptive group testing schemes. This result can be summarized in the following informal theorem.

**Theorem 1 (Informal):** For both Models 1 and 2, our deterministic nonadaptive scheme can identify the set of defectives exactly with probability  $1 - \epsilon$ . The sufficient number of tests required for this is  $O(t \frac{\ln N}{\ln t} \ln \frac{N}{\epsilon})$ .

One of our construction technique is same as the scheme of [25], [35], however with a finer analysis relying on the distance properties of a linear code we are able to achieve more. We also use explicit families of Algebraic-Geometric codes to obtain the same set of parameters. One of the main

contribution is to show a general way to establish a property for almost all  $t$ -tuples of elements from a set based on the mean pairwise statistics of the set.

In Section II, we provide the necessary definitions and preliminaries. The relation of group testing parameters of Model 1 with constant-weight codes is provided in Section III. In Section IV we establish the connection between the parameters of a weakly separated design and the average distance of a constant-weight code. In Section V we discuss our construction schemes (including one that relies on Algebraic-Geometric codes) that work for both of Models 1 and 2.

## II. BASIC DEFINITIONS AND PROPERTIES

A vector is denoted by bold lowercase letters, such as  $\mathbf{x}$ , and the  $i$ th entry of the vector  $\mathbf{x}$  is denoted by  $x_i$ . The Hamming distance between two vectors is denoted by  $d_H(\cdot, \cdot)$ . The *support* of a vector  $\mathbf{x}$  is the set of coordinates where the vector has nonzero entries. It is denoted by  $\text{supp}(\mathbf{x})$ . We use the usual set terminology, where a set  $A$  contains  $B$  if  $B \subseteq A$ . Also, below  $[n]$  denotes  $\{1, 2, \dots, n\}$ .

First of all, we define the following two models for *defectives*.

*Definition 1 (Random models of defectives):* In the random defective *Model 1*, among a set of  $N$  elements, each element is independently defective with probability  $\frac{t}{N}$ . In the random defective *Model 2*, each subset of cardinality  $t$  of a set of  $N$  elements has equal probability  $\binom{N}{t}^{-1}$  of being the defective set.

### A. Disjunct matrices

The following definition of disjunct matrices is standard and can be found in [12, Ch. 7].

*Definition 2:* An  $M \times N$  binary matrix  $A$  is called  $t$ -disjunct if the support of any column is not contained in the union of the supports of any other  $t$  columns.

It is not very difficult to see that a  $t$ -disjunct matrix gives a group testing scheme that identifies any defective set up to size  $t$ . On the other hand any group testing scheme that identifies any defective set up to size  $t$  must be a  $(t-1)$ -disjunct matrix. The definition of disjunct matrix can be restated as follows: a matrix is  $t$ -disjunct if any  $t+1$  columns indexed by  $i_1, \dots, i_{t+1}$  of the matrix form a sub matrix which must have a row that has exactly one 1 in the  $i_j$ th position and zeros in the other positions, for  $j = 1, \dots, t+1$ .

To a great advantage, disjunct matrices allow for a simple identification algorithm that runs in time  $O(Nt)$ , as we see below.

### B. Disjunct decoding

Given the test results  $\mathbf{y} \in \{0, 1\}^M$ , we use the following recovery algorithm to find the defectives. Suppose,  $A$  is the test matrix and  $\alpha^{(j)} \in \{0, 1\}^N, j = 1, \dots, M$  denotes the  $j$ th row of  $A$ . The recovery algorithm simply outputs

$$[N] \setminus \cup_{j: y_j=0} \text{supp}(\alpha^{(j)})$$

as the set of defectives [12, Ch. 7].

Note that, irrespective of the testing matrix, this algorithm will always output a set that contains all the defective elements. Moreover, if the testing matrix is disjunct, then the output is exactly equal to the set of defectives. We have the following simple proposition.

*Proposition 1:* Suppose, the set of defectives is  $S \subseteq [N]$ . Let  $\mathbf{a}^{(k)}$  denote the  $k$ th column of the test matrix  $A$ . Then the disjunct decoding algorithm recovers the defectives exactly if  $\cup_{j \in S} \text{supp}(\mathbf{a}^{(j)})$  does not contain the support of  $\mathbf{a}^{(i)}$  for all  $i \in [N] \setminus S$ .

### C. Almost disjunct matrices

Below we define a *relaxed* form of disjunct matrices. This definition appeared very closely in [29], [46] and exactly in [27].

*Definition 3:* For any  $\epsilon > 0$ , an  $M \times N$  matrix  $A$  is called  $(t, \epsilon)$ -disjunct if the set of  $t$ -tuple of columns (of size  $\binom{N}{t}$ ) has a subset  $\mathcal{B}$  of size at least  $(1 - \epsilon)\binom{N}{t}$  with the following property: for all  $J \in \mathcal{B}$ ,  $\cup_{\kappa \in J} \text{supp}(\kappa)$  does not contain support of any column  $\nu \notin J$ .

In other words, the union of supports of a randomly and uniformly chosen set of  $t$  columns from a  $(t, \epsilon)$ -disjunct matrix does not contain the support of any other column with probability at least  $1 - \epsilon$ . It is clear that for  $\epsilon = 0$ , the  $(t, \epsilon)$ -disjunct matrices are same as  $t$ -disjunct matrices.

It is easy to see the following fact.

*Proposition 2 (Model 2):* A  $(t, \epsilon)$ -disjunct matrix gives a group testing scheme that can identify all but at most a fraction  $\epsilon > 0$  of all possible defective configurations of size  $t$ .

### D. Constant-weight codes

A binary  $(M, N, d)$  code  $\mathcal{C}$  is a set of size  $N$  consisting of  $\{0, 1\}$ -vectors of length  $M$ . Here  $d$  is the largest integer such that any two vectors (codewords) of  $\mathcal{C}$  are at least Hamming distance  $d$  apart.  $d$  is called the *minimum distance* (or *distance*) of  $\mathcal{C}$ . If all the codewords of  $\mathcal{C}$  have Hamming weight  $w$ , then it is called a constant-weight code. In that case we write  $\mathcal{C}$  is an  $(M, N, d, w)$ -constant-weight binary code.

Constant-weight codes can give constructions of group testing schemes. One just arranges the codewords as the columns of the test matrix. Kautz and Singleton proved the following in [25].

*Proposition 3:* An  $(M, N, d, w)$ -constant-weight binary code provides a  $t$ -disjunct matrix where,  $t = \left\lfloor \frac{w-1}{w-d/2} \right\rfloor$ .

*Proof:* The intersection of supports of any two columns has size at most  $w - d/2$ . Hence if  $w > t(w - d/2)$ , support of any column will not be contained in the union of supports of any  $t$  other columns. ■

Extensions of Prop. 3 are our main results. To do that we need to define the *average distance*  $D$  of a code  $\mathcal{C}$ :

$$D(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \min_{\mathbf{x} \in \mathcal{C}} \sum_{\mathbf{y} \in \mathcal{C}} d_H(\mathbf{x}, \mathbf{y}).$$

Here  $d_H(\mathbf{x}, \mathbf{y})$  denotes the Hamming distance between  $\mathbf{x}$  and  $\mathbf{y}$ . Also define the second-moment of distance distribution:

$$D_2(\mathcal{C}) = \frac{1}{|\mathcal{C}|^2} \sum_{\mathbf{x}, \mathbf{y} \in \mathcal{C}} d_H(\mathbf{x}, \mathbf{y})^2.$$

### III. MODEL 1: INDEPENDENT DEFECTIVES - TEST MATRICES FROM CONSTANT-WEIGHT CODES

In this section, we consider the independent failure model (Model 1) and show how the minimum and average distances of a constant-weight binary code contribute to a nonadaptive group testing scheme. Recall, in this model we assume that among  $N$  items, each is defective with a probability  $\frac{t}{N}$ . The main result of this section is the following theorem.

*Theorem 2 (Model 1):* Suppose, we have a constant-weight binary code  $\mathcal{C}$  of size  $N$ , minimum distance  $d$  and average distance  $D$  such that every codeword has length  $M$  and weight  $w$ . The test matrix obtained from the code exactly identifies all the defective items (chosen according to Model 1) with probability at least  $1 - \epsilon$  (over the probability space of Model 1) if

$$w - \frac{d}{2} \leq \frac{3(w - t(w - D/2))^2}{2(2t(w - D/2) + w) \ln \frac{N}{\epsilon}}. \quad (1)$$

We will need the help of the following lemma to prove the theorem. Note, from Prop. 1, the disjunct-recovery algorithm will be successful if the union of supports of the columns corresponding to the defectives does not contain the support of any other columns. Suppose the testing matrix is constructed from an  $(M, N, d, w)$ -constant-weight code  $\mathcal{C}$  (each column is a codeword). Let

$$\mathcal{C} = \{c_1, c_2, \dots, c_N\}.$$

Moreover, assume  $X_j \in \{0, 1\}$  is the indicator Bernoulli( $t/N$ ) random variable that denotes whether the  $j$ th element is defective or not.

*Lemma 1:* Suppose, for all  $i \in [N]$ , we have

$$\sum_{j=1, j \neq i}^N X_j \left( w - \frac{d_H(c_i, c_j)}{2} \right) < w.$$

Then the disjunct-recovery algorithm will exactly identify the defective elements.

*Proof:* The lemma directly follows from Prop. 1 and the fact that for any  $i, j$ ,  $w - \frac{d_H(c_i, c_j)}{2}$  is nonnegative. Suppose  $S \subseteq [N]$  be the random set of defectives. The disjunct-recovery algorithm will be successful when for all  $i \in [N] \setminus S$ ,

$$\sum_{j \in S} \left( w - \frac{d_H(c_i, c_j)}{2} \right) < w.$$

Hence the condition of the lemma is sufficient for success. ■

Now we are ready to prove Thm. 2.

*Proof of Thm. 2:* First of all, by union bound,

$$\begin{aligned} \Pr \left( \exists i \in [N] : \sum_{j=1, j \neq i}^N X_j \left( w - \frac{d_H(c_i, c_j)}{2} \right) \geq w \right) \\ \leq \sum_i \Pr \left( \sum_{j=1, j \neq i}^N X_j \left( w - \frac{d_H(c_i, c_j)}{2} \right) \geq w \right). \end{aligned}$$

For a fixed  $i$ , we would want to upper bound the probability above in the right hand side under the summation. Assume,  $w - \frac{d_H(c_i, c_j)}{2} = a_j$ . Notice,  $a_j X_j - \mathbb{E}(a_j X_j) \leq a_j(1 - t/N) \leq$

$(1 - t/N)(w - d/2)$  and  $\sum_{j \neq i} \mathbb{E}(a_j X_j - a_j t/N)^2 = \frac{t}{N} \left( 1 - \frac{t}{N} \right) \sum_{j \neq i} a_j^2$ . We have,

$$\begin{aligned} \Pr \left( \sum_{j=1, j \neq i}^N X_j \left( w - \frac{d_H(c_i, c_j)}{2} \right) \geq w \right) \\ = \Pr \left( \sum_{j=1, j \neq i}^N (X_j - t/N) a_j \geq w - \frac{t}{N} \sum_{j \neq i} a_j \right) \\ \stackrel{(a)}{\leq} \Pr \left( \sum_{j=1}^N (X_j - t/N) a_j \geq w - \frac{t}{N} \sum_{j=1}^N a_j \right), \end{aligned}$$

where (a) is true as the event within the probability in second line implies the event in the third line.

Now, we can use the classical Bernstein concentration inequality (see the version we use in [32, Thm. 2.7]), to have,

$$\begin{aligned} -\ln \Pr \left( \sum_{j=1, j \neq i}^N X_j \left( w - \frac{d_H(c_i, c_j)}{2} \right) \geq w \right) \\ \geq \frac{\left( w - \frac{t}{N} \sum_j a_j \right)^2}{2 \left( 1 - \frac{t}{N} \right) \left( \frac{t}{N} \sum_j a_j^2 + \frac{1}{3} \left( w - \frac{d}{2} \right) \left( w - \frac{t}{N} \sum_j a_j \right) \right)} \\ \geq \frac{\left( w - \frac{t}{N} \sum_j a_j \right)^2}{2 \left( \frac{t}{N} \sum_j a_j^2 + \frac{1}{3} \left( w - \frac{d}{2} \right) \left( w - \frac{t}{N} \sum_j a_j \right) \right)} \\ \geq \frac{\left( w - \frac{t}{N} \sum_j a_j \right)^2}{2 \left( \frac{t}{N} \sum_j a_j \left( w - \frac{d}{2} \right) + \frac{1}{3} \left( w - \frac{d}{2} \right) \left( w - \frac{t}{N} \sum_j a_j \right) \right)} \\ \geq \frac{3 \left( w - \frac{t}{N} \sum_j a_j \right)^2}{2 \left( w - \frac{d}{2} \right) \left( \frac{2t}{N} \sum_j a_j + w \right)} \\ \stackrel{(b)}{\geq} \frac{3 \left( w - t(w - D/2) \right)^2}{2 \left( w - \frac{d}{2} \right) \left( 2t(w - D/2) + w \right)}, \end{aligned}$$

and (b) follows because the exponent above is an increasing function of  $\sum_j a_j$  and  $\frac{1}{N} \sum_j a_j = w - \frac{1}{2N} \sum_j d_H(c_i, c_j) \leq w - \frac{D}{2}$ . Now using union bound, we deduce that the test matrix will successfully identify the defective elements exactly with probability  $1 - \epsilon$  if

$$\frac{3 \left( w - t(w - D/2) \right)^2}{2 \left( w - \frac{d}{2} \right) \left( 2t(w - D/2) + w \right)} \geq \ln \frac{N}{\epsilon},$$

which proves the theorem. ■

Similar result can be obtained for Model 2. However, because of the dependence among the random choice of defectives we need to use concentration inequalities for sampling without replacements.

### IV. MODEL 2: $(t, \epsilon)$ -DISJUNCT MATRICES FROM CONSTANT-WEIGHT CODES

Our main result of this section is the following.



*Theorem 3 (Model 2):* Suppose, we have a constant-weight binary code  $\mathcal{C}$  of size  $N$ , minimum distance  $d$  and average distance  $D$  such that every codeword has length  $M$  and weight  $w$ . The test matrix obtained from the code is  $(t, \epsilon)$ -disjunct for the largest  $t$  such that,

$$d \geq D - \frac{3(w - t(w - D/2))^2}{\ln \frac{N}{\epsilon} (2t(w - D/2) + w)},$$

holds.

One can compare the results of Prop. 3 and Theorem 3 to see the improvement achieved as we relax the definition of disjunct matrices. Indeed, Theorem 3 implies,

$$t \leq \frac{w - \sqrt{\frac{1}{3}(D - d) \ln \frac{N}{\epsilon} (2t(w - D/2) + w)}}{w - D/2},$$

as opposed to  $t \leq \frac{w-1}{w-d/2}$  from Prop. 3. This will lead to the final improvement on the parameters of Porat-Rothschild construction [35], as we will see in Section V.

#### A. Proof of Theorem 3

This section is dedicated to the proof of Theorem 3. Suppose, we have a constant-weight binary code  $\mathcal{C}$  of size  $N$  and minimum distance  $d$  such that every codeword has length  $M$  and weight  $w$ . Let the average distance of the code be  $D$ . Note that this code is fixed: we will prove the almost-disjunctness property of this code.

Let us now choose  $t$  codewords randomly and uniformly from all possible  $\binom{N}{t}$  choices. Let the randomly chosen codewords be  $\{c_1, c_2, \dots, c_t\}$ . In what follows, we adapt the proof of Prop. 3 in a probabilistic setting.

Assume we call the random set of defectives as  $S$ . For  $l \in [N] \setminus S$ , define the random variables  $Z^l = \sum_{j=1}^t \left( w - \frac{d_H(c_l, c_j)}{2} \right)$ . Clearly,  $Z^l$  is the maximum possible size of the portion of the support of  $c_l$  that is common to at least one of  $c_j, j = 1, \dots, t$ . Note that the size of support of  $c_l$  is  $w$ . Hence, as we have seen in the proof of Prop. 3, if  $Z^l$  is less than  $w$  for all  $l$ 's that are not part of the defective set, then the disjunct decoding algorithm will be successful. Therefore, we aim to find the probability  $\Pr(\exists l \in [N] \setminus S : Z^l \geq w)$  and show it to be bounded above by  $\epsilon$  under the condition of the theorem.

As the variable  $Z^l$ 's are identically distributed, using union bound,

$$\Pr(\exists l \in [N] \setminus S : Z^l \geq w) \leq (N - t) \Pr(Z^l \geq w),$$

where  $l$  can now assumed to be uniformly distributed in  $[N] \setminus S$ . In the following, we will find an upper bound on  $\Pr(Z^l \geq w)$ .

In [30], an upper bound on  $\Pr(Z^l \geq w)$  was found by Azuma's inequality. It turns out that by using a trick from Hoeffding [20], and using the Bernstein inequality we can achieve a tighter bound. First note that,

$$Z^l = \sum_{j=1}^t \left( w - \frac{d_H(c_l, c_j)}{2} \right),$$

where,  $c_1, \dots, c_t, c_l$  are randomly and uniformly chosen  $(t+1)$  codewords from all possible  $\binom{N}{t+1}$  choices.

Given,  $c_l$ , the other codewords are randomly sampled from the code without replacement. It follows from [20, Theorem 4], for any real number  $s$  that,

$$\mathbb{E}(e^{sZ^l} | c_l) \leq \mathbb{E}\left(e^{s \sum_{j=1}^t \left( w - \frac{d_H(c_l, c_j)}{2} \right)} | c_l\right),$$

where  $x_1, \dots, x_t$  are codewords randomly and uniformly sampled from the code with replacement. Therefore,

$$\mathbb{E}(e^{sZ^l}) \leq \mathbb{E}\left(e^{s \sum_{j=1}^t \left( w - \frac{d_H(c_l, x_j)}{2} \right)}\right),$$

where  $c_l$  is a randomly and uniformly chosen codeword and  $x_1, \dots, x_t$  are codewords randomly and uniformly sampled from the code  $\mathcal{C} \setminus \{c_l\}$  with replacement.

Therefore, for any  $s > 0$ , using Markov inequality,

$$\Pr(Z^l \geq w) \leq \mathbb{E}e^{-sw} \left( e^{s \sum_{j=1}^t Y_j} \right),$$

where,  $Y_i \equiv w - \frac{d_H(c_l, x_i)}{2}, i = 1, \dots, t$ , are independent random variables with,

$$\mathbb{E}Y_i \leq w - \frac{D}{2}.$$

and

$$\mathbb{E}Y_i^2 \leq \left( w - \frac{D}{2} \right) \left( w - \frac{d}{2} \right),$$

since  $Y_i$  is a nonnegative random variable. Now, since  $Y_i$ 's are all independent, we can use [32, Thm. 2.7] (or its method of proof) again, to upper bound large deviation for the sum  $Z^l$ . Indeed, we must have,

$$\Pr(Z^l \geq w) \leq \exp\left(-\frac{(w - \sum_{i=1}^t \mathbb{E}Y_i)^2}{\mathcal{A}}\right),$$

where,

$$\begin{aligned} \mathcal{A} &= 2 \sum_{i=1}^t (\mathbb{E}Y_i^2 - (\mathbb{E}Y_i)^2) \\ &\quad + \frac{2}{3} \left( w - \sum_{i=1}^t \mathbb{E}Y_i \right) \left( w - \frac{d}{2} - w + \frac{D}{2} \right) \\ &\leq 2t \left( w - \frac{D}{2} \right) \left( w - \frac{d}{2} \right) - 2t \left( w - \frac{D}{2} \right)^2 \\ &\quad + \frac{1}{3} \left( w - t \left( w - \frac{D}{2} \right) \right) (D - d) \\ &= t \left( w - \frac{D}{2} \right) (D - d) + \frac{1}{3} \left( w - t \left( w - \frac{D}{2} \right) \right) (D - d). \end{aligned}$$

Hence, we have,

$$\Pr(Z^l \geq w) \leq \exp\left(-\frac{3(w - t(w - \frac{D}{2}))^2}{(2t(w - \frac{D}{2}) + w)(D - d)}\right).$$

Now using union bound, we deduce that the test matrix will successfully identify the defective elements exactly with probability  $1 - \epsilon$  if

$$\frac{3(w - t(w - D/2))^2}{(D - d)(2t(w - D/2) + w)} \geq \ln \frac{N}{\epsilon},$$

which proves the theorem.

### B. Higher order statistics of distance distribution

We get slightly tighter bounds in both the Theorems 2 and 3, if higher order than only the average distance of the codes have been considered. Indeed in both of the main theorems we have used the inequality,

$$\frac{1}{|\mathcal{C}|^2} \sum_{\mathbf{y}, \mathbf{x} \in \mathcal{C}} \left( w - \frac{d_H(\mathbf{x}, \mathbf{y})}{2} \right)^2 \leq \left( w - \frac{d}{2} \right) \frac{1}{|\mathcal{C}|^2} \sum_{\mathbf{y}, \mathbf{x} \in \mathcal{C}} \left( w - \frac{d_H(\mathbf{x}, \mathbf{y})}{2} \right),$$

since  $w - \frac{d_H(\mathbf{x}, \mathbf{y})}{2}$  is always nonnegative. However both of the theorems could be rephrased in terms of the second-moment of the distance distribution. For example, Theorem 3 can be restated with slightly stronger result.

**Theorem 4:** Suppose, we have a constant-weight  $(M, N, d, w)$  binary code  $\mathcal{C}$  with average distance  $D$  and the second-moment of the distance distribution  $D_2$ . The test matrix obtained from the code is  $(t, \epsilon)$ -disjunct for the largest  $t$  such that,

$$d \geq D + \frac{3t(D_2 - D^2)}{2(w - t(w - D/2))} - \frac{3(w - t(w - D/2))}{\ln \frac{N}{\epsilon}}, \quad (2)$$

holds.

We omit the proof of this theorem as it is exactly same as the proof of Theorem 3.

However, it turns out (in the next section) that our results, that rely only on the average distance, are sufficient to give near-optimal performance in group testing schemes in terms of the number of tests. In particular, use of (??) in conjunction with the construction of constant-weight codes below, instead of Theorem 3, leads to improvement only on the constant terms.

## V. CONSTRUCTION

### A. Discussions

As we have seen in Section II, constant-weight codes can be used to produce disjunct matrices. Kautz and Singleton [25] gives a construction of constant-weight codes that results in good disjunct matrices. In their construction, they start with a Reed-Solomon (RS) code, a  $q$ -ary error-correcting code of length  $q-1$ . For a detailed discussion of RS codes we refer the reader to the standard textbooks of coding theory [28], [36]. Next they replace the  $q$ -ary symbols in the codewords by unit weight binary vectors of length  $q$ . The mapping from  $q$ -ary symbols to length- $q$  unit weight binary vectors is bijective: i.e., it is  $0 \rightarrow 100 \dots 0; 1 \rightarrow 010 \dots 0; \dots; q-1 \rightarrow 0 \dots 01$ . We refer to this mapping as  $\phi$ . As a result, one obtains a set of binary vectors of length  $q(q-1)$  and constant-weight  $q$ . The size of the resulting binary code is same as the size of the RS code, and the distance of the binary code is twice that of the distance of the RS code.

For a  $q$ -ary RS code of size  $N$  and length  $q-1$ , the minimum distance is  $q-1-\log_q N+1 = q-\log_q N$ . Hence, the Kautz-Singleton construction is a constant-weight code with length  $M = q(q-1)$ , weight  $w = q-1$ , size  $N$  and distance

$2(q-\log_q N)$ . Therefore, from Prop. 3, we have a  $t$ -disjunct matrix with,

$$t = \frac{q-1-1}{q-1-q+\log_q N} = \frac{q-2}{\log_q N-1} \\ \approx \frac{q \log q}{\log N} \approx \frac{\sqrt{M} \log M}{2 \log N}.$$

On the other hand, note that, the average distance of the RS code is  $(q-1)(1-1/q)$ . Hence the average distance of the resulting constant-weight code from Kautz-Singleton construction will be

$$D = \frac{2(q-1)^2}{q}.$$

Now, substituting these values in Theorem 3, we have a  $(t, \epsilon)$  disjunct matrix, where,

$$2(q-\log_q N) \\ \geq 2 \frac{(q-1)^2}{q} - \frac{3(q-1-t(q-1-\frac{(q-1)^2}{q}))^2}{(2t(q-1-\frac{(q-1)^2}{q})+q-1) \ln \frac{N}{\epsilon}} \\ = \frac{2(q-1)^2}{q} - \frac{3(q-1)(1-t/q)^2}{(1+2t/q) \ln \frac{N}{\epsilon}}.$$

This basically restricts  $t$  to be about  $O(\sqrt{M})$  (since,  $1-t/q$  must be nonnegative). Hence, Theorem 3 does not obtain any meaningful improvement from the Kautz-Singleton construction in the asymptotics except in special cases.

There are two places where the Kautz-Singleton construction can be modified: 1) instead of Reed-Solomon code one can use any other  $q$ -ary code of different length, and 2) instead of the mapping  $\phi$  any binary constant-weight code of size  $q$  might have been used. For a general discussion we refer the reader to [12, §7.4]. In the recent work [35], the mapping  $\phi$  is kept the same, while the RS code has been changed to a  $q$ -ary code that achieve the Gilbert-Varshamov bound [28], [36].

In our construction of disjunct matrices we use the Kautz-Singleton construction and instead of Reed-Solomon code either 1) follow the footsteps of [35] to use a Gilbert-Varshamov code or 2) use Algebraic-Geometric codes. We exploit some property of the resulting scheme (namely, the average distance) and do a finer analysis that was absent from the previous works such as [35].

### B. $q$ -ary Gilbert-Varshamov construction

Next, we construct a linear  $q$ -ary code of size  $N$ , length  $M_q$  and minimum distance  $d_q$  that achieves the Gilbert-Varshamov (GV) bound [28], [36]. We describe the bound in Appendix A.

Porat and Rothschild [35] show that it is possible to construct in time  $O(M_q N)$  a  $q$ -ary code that achieves the GV bound. To have such construction, they exploit the following well-known fact: a  $q$ -ary linear code with random generator matrix achieves the GV bound with high probability [36]. To have an explicit construction of such codes, a derandomization method known as the method of conditional expectation [1] is used. In this method, the entries of the generator matrix of the code are chosen one-by-one so that the minimum distance of

the resulting code does not go below the value prescribed by Eq. (??). For a detail description of the procedure, see [35].

Using the GV code construction of Porat and Rothschild and plugging it in the Kautz-Singleton construction above, we have the following proposition.

*Proposition 4:* Let  $s \leq q$ . There exists a polynomial time constructible family of  $(M, N, 2M/q(1 - 1/s), M/q)$ -constant-weight binary code that satisfy,

$$M/q \leq \frac{s \ln N}{\ln(q/s) - 1}. \quad (3)$$

Although the proof of the above proposition is essentially in Porat and Rothschild [35], we have a cleaner proof that we include in Appendix A for completeness.

However, we are also concerned with the average distance of the code. Indeed, we have the following proposition.

*Proposition 5:* The average distance of the code constructed in Prop. 4 is

$$D = \frac{2M}{q}(1 - 1/q).$$

*Proof:* For Prop. 4 we have followed the Kautz-Singleton construction. We take a linear  $q$ -ary code  $C'$  of length  $M_q \triangleq \frac{M}{q}$ , size  $N$  and minimum distance  $d_q \triangleq \frac{d}{2}$ . Each  $q$ -ary symbol in the codewords is then replaced with a binary indicator vector of length  $q$  (i.e., the binary vector whose all entries are zero but one entry, which is 1) according to the map  $\phi$ . As a result we have a binary code  $C$  of length  $M$  and size  $N$ . The minimum distance of the code is  $d$  and the codewords are of constant-weight  $w = M_q = \frac{M}{q}$ . The average distance of this code is twice the average distance of the  $q$ -ary code. As  $C'$  is linear (assuming it has no all-zero coordinate), it has average distance equal to

$$\begin{aligned} \frac{1}{N} \sum_{j=1}^{M_q} j A_j &= \frac{N}{N} \sum_{j=0}^{M_q} j \binom{M_q}{j} (1 - 1/q)^j (1/q)^{M_q-j} \\ &= M_q(1 - 1/q), \end{aligned}$$

where  $A_j$  is the number of codewords of weight  $j$  in  $C'$ . Here we use the fact that the average of the distance between any two randomly chosen codewords of a nontrivial linear code is equal to that of a binomial random variable [28]. Hence the constant-weight code  $C$  has average distance  $D = 2M_q(1 - 1/q)$ . ■

### C. Constructions for Model 1

We follow the Kautz-Singleton code construction. Suppose, we have a  $(M, N, d, M/q)$ -constant-weight code that satisfies Prop. 4 and 5. Hence, average distance  $D = \frac{2M}{q}(1 - 1/q)$ . The resulting test matrix will satisfy the condition of Thm. 2 when,

$$d/2 \geq M/q - \frac{3\left(M/q - t(M/q - M/q(1 - 1/q))\right)^2}{2\left(2t(M/q - M/q(1 - 1/q)) + M/q\right) \ln \frac{N}{\epsilon}} \quad (4)$$

or when,

$$d/2 \geq M/q - \frac{3M/q\left(1 - t/q\right)^2}{2\left(2t/q + 1\right) \ln \frac{N}{\epsilon}}. \quad (5)$$

Hence a sufficient condition is to choose the constant-weight code such that,

$$d \geq \frac{2M}{q} \left(1 - \frac{3\left(1 - t/q\right)^2}{2\left(2t/q + 1\right) \ln \frac{N}{\epsilon}}\right).$$

We can take  $q$  to be the smallest power of prime that is greater than  $2t$ . Which will make the sufficient condition look like,

$$d \geq \frac{2M}{q} \left(1 - \frac{3}{16 \ln \frac{N}{\epsilon}}\right).$$

However, according to Prop. 4, such code can be explicitly constructed with,

$$M/q \leq \frac{16/3 \ln \frac{N}{\epsilon} \ln N}{\ln(3t/(16 \ln \frac{N}{\epsilon})) - 1}. \quad (6)$$

Hence, the sufficient number of tests is  $M = \frac{6t}{\ln t} \ln N \ln \frac{N}{\epsilon}$ .

### D. Construction of almost disjoint matrix: Model 2

We again follow the above code construction and choose  $q$  to be a power of a prime number. With proper parameters we can have a disjoint matrix with the following property.

*Theorem 5:* It is possible to explicitly construct a  $(t, \epsilon)$ -disjoint matrix of size  $M \times N$  where

$$M = O\left(\frac{t}{\log t} \log N \log \frac{N}{\epsilon}\right).$$

*Proof:* We follow the Kautz-Singleton code construction as earlier. That is we have a  $(M, N, d, M/q)$ -constant-weight code that satisfies Prop. 4 and 5. Hence, average distance  $D = \frac{2M}{q}(1 - 1/q)$ . The resulting matrix will be  $(t, \epsilon)$ -disjoint if the condition of Theorem 3 is satisfied, i.e.,

$$\begin{aligned} d &\geq \frac{2M}{q}(1 - 1/q) \\ &\quad - \frac{3\left(M/q - t(M/q - M/q(1 - 1/q))\right)^2}{\left(2t(M/q - M/q(1 - 1/q)) + M/q\right) \ln \frac{N}{\epsilon}}, \end{aligned}$$

or when,

$$d \geq \frac{2M}{q}(1 - 1/q) - \frac{3M/q\left(1 - t/q\right)^2}{\left(2t/q + 1\right) \ln \frac{N}{\epsilon}}. \quad (7)$$

Hence a sufficient condition is to choose the constant-weight code such that,

$$d \geq \frac{2M}{q} \left(1 - \frac{1}{q} - \frac{3\left(1 - t/q\right)^2}{2\left(2t/q + 1\right) \ln \frac{N}{\epsilon}}\right).$$

Since the requirement of above sufficient condition is slightly weaker than that of (??), we can still choose  $q$  to be a

smallest power of prime that is greater than  $2t$ , and follow the calculations for Model 1, to obtain total number of tests  $M = O\left(\frac{t}{\ln t} \ln N \ln \frac{N}{\epsilon}\right)$ . ■

It is clear from Prop. 2 that a  $(t, \epsilon)$  disjunct matrix is equivalent to a group testing scheme. Hence, as a consequence of Theorem 5, we will be able to construct a testing scheme with  $O\left(\frac{t}{\log t} \ln N \ln \frac{N}{\epsilon}\right)$  tests. Whenever the defect-model is such that all the possible defective sets of size  $t$  are equally likely and there are no more than  $t$  defective elements, the above group testing scheme will be successful with probability at least  $1 - \epsilon$ .

Note that, if  $t$  is proportional to any positive power of  $N$ , then  $\log N$  and  $\log t$  are of same order. Hence it will be possible to have the above testing scheme with  $O(t \log \frac{N}{\epsilon})$  tests, for any  $\epsilon > 0$ .

### E. Constructions based on Algebraic-Geometric codes

Now, instead of using the Porat-Rothschild construction of GV codes, we can use the Algebraic-Geometric (AG) code construction of Tsfasman, Vlăduț and Zink [41]. In particular, we can base our construction on the García-Stichtenoth Tower of function field over  $\mathbb{F}_q$  [42, Sec. 3.4.3].

Assume,  $q = r^2$ , where  $r$  is any integer. For any even number  $n$ , there is a family of modular curves with genus  $g_n = (r^{n/2} - 1)^2$  with number of points given by  $M_q \geq r^{n+1} - r^n + 1$  (see, [42, Theorem 3.4.44]). Now, using Corollary 4.1.14 of [42], we conclude that it is possible to construct families of linear code of length  $M_q$ , size  $N$  and minimum distance  $d_q$ , where,

$$M_q \geq r^{n+1} - r^n + 1,$$

and

$$\log_q N = M_q - d_q - g_n + 1.$$

Hence, we obtain families of linear code such that,

$$\frac{\log_q N}{M_q} \geq 1 - \frac{d_q}{M_q} - \frac{1}{\sqrt{q} - 1}. \quad (8)$$

Now, using the Kautz-Singleton mechanism of converting this to a binary code, we obtain an  $(M, N, d, w)$  constant weight code, where

$$M = qM_q; d = 2d_q; w = M_q = M/q,$$

and,

$$d \geq \frac{2M}{q} \left(1 - \frac{q \log_q N}{M} - \frac{1}{\sqrt{q} - 1}\right). \quad (9)$$

Since, the AG code is a linear code, we can calculate the the average distance of the above constant-weight code as in Proposition 5. Indeed, the average distance  $D = \frac{2M}{q} \left(1 - \frac{1}{q}\right)$ .

For this family of codes, we can also calculate the second-moment of the distance distribution<sup>1</sup>, that allows us to use Theorem 4. To be consistent of the rest of the paper, we rely on only the average distance, and use Theorem 3 instead.

<sup>1</sup>It turns out that  $D_2 = D^2 + \frac{4M}{q^2} \left(1 - \frac{1}{q}\right)$ .

Substituting the values of  $D, w$  in Theorem 3, we obtain the following. If

$$d \geq \frac{2M}{q} \left(1 - \frac{1}{q} - \frac{3 \left(1 - t/q\right)^2}{2 \left(2t/q + 1\right) \ln \frac{N}{\epsilon}}\right), \quad (10)$$

then the construction is  $(t, \epsilon)$ -almost disjunct. Comparing (??) and (??), we claim that, our construction is  $(t, \epsilon)$ -almost disjunct as long as,

$$\frac{q \log_q N}{M} + \frac{1}{\sqrt{q} - 1} \leq \frac{1}{q} + \frac{3 \left(1 - t/q\right)^2}{2 \left(2t/q + 1\right) \ln \frac{N}{\epsilon}}. \quad (11)$$

Now assuming  $q$  to be the smallest power of 2 greater than  $2t$ , we see that the above condition is satisfied when,

$$M \geq \frac{16t \ln N}{\ln 2t} \ln \frac{N}{\epsilon}.$$

We should note that construction for Model 1 can be done in the exact same way to obtain the same parameters.

*Remark 1: (The traditional argument (Prop. 3) with Algebraic-Geometric Codes)* Note that, one could use AG codes in conjunction with Prop. 3 to obtain disjunct matrices. However such a construction results in highly suboptimal number of rows (tests). Indeed, substituting Eq. (??) in Prop. 3, we have a  $t$ -disjunct matrix with,

$$t = \frac{1}{\frac{q \log_q N}{M} + \frac{1}{\sqrt{q} - 1}} \Rightarrow \frac{q \log_q N}{M} = \frac{1}{t} - \frac{1}{\sqrt{q} - 1}.$$

Hence, to get anything nontrivial we must have  $q \geq t^2$ , which results in  $M = \Omega(t^3 \log N / \log t)$ . This is quite bad compared to the optimal constructions that give disjunct matrices with  $O(t^2 \log N)$  rows. It is interesting that by using our average distance based arguments we are able to get rid of such suboptimality with AG codes. Intuitively, while the range of minimum distance of the constant-weight codes obtained from the AG codes is not sufficient for optimal results, the combination of average distance and minimum distance for these codes indeed belongs to the best possible region.

## VI. CONCLUSION

In this work we show that it is possible to construct non-adaptive group testing schemes with small number of tests that identify a uniformly chosen random defective configuration with high probability. To construct a  $t$ -disjunct matrix one starts with the simple relation between the minimum distance  $d$  of a constant  $w$ -weight code and  $t$ . This is an example of a scenario where a pairwise property (i.e., distance) of the elements of a set is translated into a property of  $t$ -tuples.

Our method of analysis provides a general way to prove that a property holds for almost all  $t$ -tuples of elements from a set based on the mean pairwise statistics of the set. Our method might be useful in many areas of applied combinatorics, such as digital fingerprinting or design of key-distribution schemes, where such a translation is evident. With this method potential new results may be obtained for the cases of cover-free codes [14], [25], [40], traceability and frameproof codes [10], [39].



## APPENDIX

## A. Gilbert-Varshamov bound and proof of Prop. 4

**Lemma 2 (Gilbert-Varshamov Bound):** There exists an  $(m, N, d)_q$ -code such that,

$$N \geq \frac{q^m}{\sum_{i=0}^{d-1} \binom{m}{i} (q-1)^i}. \quad (12)$$

**Corollary 1:** Suppose  $X$  is a Binomial( $m, 1 - \frac{1}{q}$ ) random variable. There exists an  $(m, N, d)_q$ -code such that,

$$N \geq \frac{1}{\Pr(X \leq d)}.$$

**Lemma 3:** Suppose  $X$  is a Binomial( $m, 1 - \frac{1}{q}$ ) random variable. Then, for all  $s < q$ ,

$$\Pr\left(X \leq m\left(1 - \frac{1}{s}\right)\right) \leq e^{-mD(1/s||1/q)}, \quad (13)$$

where  $D(p||p') = p \ln(p/p') + (1-p) \ln((1-p)/(1-p'))$ .

**Theorem 6:** Let  $s < q$ . For the  $(m, N, m(1 - 1/s))_q$ -code that achieves the Gilbert-Varshamov bound, we have

$$m \leq \frac{s \ln N}{\ln(q/s) - 1}. \quad (14)$$

**Proof:** This theorem follows from corollary 1 and lemma 3. Note that,

$$\begin{aligned} D(1/s||1/q) &= \frac{1}{s} \ln \frac{q}{s} + \left(1 - \frac{1}{s}\right) \ln \left(1 - \frac{1}{s}\right) \\ &\quad - \left(1 - \frac{1}{s}\right) \ln \left(1 - \frac{1}{q}\right) \\ &\geq \frac{1}{s} \ln \frac{q}{s} + \left(1 - \frac{1}{s}\right) \ln \left(1 - \frac{1}{s}\right) \\ &\geq \frac{1}{s} \ln \frac{q}{s} - \frac{1}{s}, \end{aligned}$$

where in the last line we have used the fact that  $x \ln x \geq x - 1$  for all  $x > 0$ . ■

Using the Kautz-Singleton construction, this implies that, there exists a polynomial time constructible family of  $(M, N, 2M/q(1 - 1/s), M/q)$ -constant-weight binary code with,

$$M/q \leq \frac{s \ln N}{\ln(q/s) - 1},$$

which is Prop. 4.

**Acknowledgements:** The author would like to thank Alexander Barg for many discussions related to the group testing problem.

## REFERENCES

- [1] N. Alon and J. H. Spencer. *The Probabilistic Method*. Wiley, New York, NY, 2004.
- [2] G. K. Atia and V. Saligrama. Boolean compressed sensing and noisy group testing. *Information Theory, IEEE Transactions on*, 58(3):1880–1901, 2012.
- [3] T. Berger, N. Mehravari, D. Towsley, and J. Wolf. Random multiple-access communication and group testing. *IEEE Transactions on Communications*, 32(7):769–779, 1984.
- [4] S. Blumenthal, S. Kumar, and M. Sobel. A symmetric binomial group-testing with three outcomes. In *Purdue Symp. Decision Procedures*, 1971.
- [5] S. Cai, M. Jahangoshahi, M. Bakshi, and S. Jaggi. Grotesque: Noisy group testing (quick and efficient). *arXiv preprint arXiv:1307.2811*, 2013.
- [6] R. Calderbank, S. Howard, and S. Jafarpour. Construction of a large class of deterministic sensing matrices that satisfy a statistical isometry property. *Selected Topics in Signal Processing, IEEE Journal of*, 4(2):358–374, 2010.
- [7] C. L. Chan, S. Jaggi, V. Saligrama, and S. Agnihotri. Non-adaptive group testing: Explicit bounds and novel algorithms. In *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*, pages 1837–1841. IEEE, 2012.
- [8] M. Cheraghchi. Noise-resilient group testing: Limitations and constructions. In *Fundamentals of Computation Theory*, pages 62–73. Springer, 2009.
- [9] M. Cheraghchi. Improved constructions for non-adaptive threshold group testing. *Algorithmica*, 67(3):384–417, 2013.
- [10] B. Chor, A. Fiat, and M. Naor. Tracing traitors. In *Advances in cryptology CRYPTO94*, pages 257–270. Springer, 1994.
- [11] R. Dorfman. The detection of defective members of large populations. *The Annals of Mathematical Statistics*, 14(4):436–440, 1943.
- [12] D. Z. Du and F. Hwang. *Combinatorial group testing and its applications, 2nd Ed.* World Scientific, 2000.
- [13] A. Dyachkov, V. Rykov, and A. Rashad. Superimposed distance codes. *PROBLEMS OF CONTROL AND INFORMATION THEORY-PROBLEMY UPRAVLENIYA I TEORII INFORMATSII*, 18(4):237–250, 1989.
- [14] A. D'yachkov, P. Vilenkin, D. Torney, and A. Macula. Families of finite sets in which no intersection of  $l$  sets is covered by the union of  $s$  others. *Journal of Combinatorial Theory, Series A*, 99(2):195–218, 2002.
- [15] A. G. D'yachkov, A. J. Macula Jr, and V. V. Rykov. New constructions of superimposed codes. *Information Theory, IEEE Transactions on*, 46(1):284–290, 2000.
- [16] A. G. D'yachkov and V. V. Rykov. Bounds on the length of disjunctive codes. *Problemy Peredachi Informatsii*, 18(3):7–13, 1982.
- [17] A. G. Dyachkov and V. V. Rykov. A survey of superimposed code theory. *Problems of Control and Information Theory*, 12(4), 1983.
- [18] A. C. Gilbert, B. Hemenway, A. Rudra, M. J. Strauss, and M. Wootters. Recovering simple signals. In *Information Theory and Applications Workshop (ITA), 2012*, pages 382–391. IEEE, 2012.
- [19] A. C. Gilbert, M. A. Iwen, and M. J. Strauss. Group testing and sparse signal recovery. In *Signals, Systems and Computers, 2008 42nd Asilomar Conference on*, pages 1059–1063. IEEE, 2008.
- [20] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.
- [21] F. Hwang. A method for detecting all defective members in a population by group testing. *Journal of the American Statistical Association*, 67(339):605–608, 1972.
- [22] F. Hwang and V. Sós. Non-adaptive hypergeometric group testing. *Studia Sci. Math. Hungar.*, 22:257–263, 1987.
- [23] F. K. Hwang. Three versions of a group testing game. *SIAM Journal on Algebraic Discrete Methods*, 5(2):145–153, 1984.
- [24] P. Indyk, H. Q. Ngo, and A. Rudra. Efficiently decodable non-adaptive group testing. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1126–1142. Society for Industrial and Applied Mathematics, 2010.
- [25] W. Kautz and R. Singleton. Nonrandom binary superimposed codes. *Information Theory, IEEE Transactions on*, 10(4):363–377, 1964.
- [26] A. J. Macula and L. J. Popyack. A group testing method for finding patterns in data. *Discrete applied mathematics*, 144(1):149–157, 2004.
- [27] A. J. Macula, V. V. Rykov, and S. Yekhanin. Trivial two-stage group testing for complexes using almost disjunct matrices. *Discrete Applied Mathematics*, 137(1):97–107, 2004.
- [28] F. J. MacWilliams and N. J. A. Sloane. *The Theory of Error-Correcting Codes*. North-Holland, 1977.
- [29] M. B. Mal'utov. The separating property of random matrices. *Mathematical Notes*, 23(1):84–91, 1978.
- [30] A. Mazumdar. On almost disjunct matrices for group testing. In *Algorithms and Computation*, pages 649–658. Springer, 2012.
- [31] A. Mazumdar and A. Barg. Sparse recovery properties of statistical RIP matrices. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pages 9–12. IEEE, 2011.
- [32] C. McDiarmid. Concentration. In *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998. [Online: [http://www.stats.ox.ac.uk/people/academic\\_staff/colin\\_mcdiarmid/?a=4139](http://www.stats.ox.ac.uk/people/academic_staff/colin_mcdiarmid/?a=4139); accessed 22-Aug-2016].

- [33] H. Q. Ngo and D.-Z. Du. A survey on combinatorial group testing algorithms with applications to dna library screening. *Discrete Mathematical Problems with Medical Applications*, 55:171–182, 2000.
- [34] H. Q. Ngo, E. Porat, and A. Rudra. Efficiently decodable error-correcting list disjunct matrices and applications. In *Automata, Languages and Programming*, pages 557–568. Springer, 2011.
- [35] E. Porat and A. Rothschild. Explicit non-adaptive combinatorial group testing schemes. In *Automata, Languages and Programming*, pages 748–759. Springer, 2008.
- [36] R. M. Roth. *Introduction to Coding Theory*. Cambridge University Press, 2006.
- [37] J. Scarlett and V. Cevher. Phase transitions in group testing. In *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 40–53. SIAM, 2016.
- [38] M. Sobel and P. A. Groll. Binomial Group-Testing With an Unknown Proportion of Defectives. *Technometrics*, 8:631–656, 1966.
- [39] J. N. Staddon, D. R. Stinson, and R. Wei. Combinatorial properties of frameproof and traceability codes. *Information Theory, IEEE Transactions on*, 47(3):1042–1049, 2001.
- [40] D. R. Stinson, R. Wei, and L. Zhu. Some new bounds for cover-free families. *Journal of Combinatorial Theory, Series A*, 90(1):224–234, 2000.
- [41] M. A. Tsfasman, S. Vlăduț, and T. Zink. Modular curves, Shimura curves, and Goppa codes, better than Varshamov-Gilbert bound. *Mathematische Nachrichten*, 109(1):21–28, 1982.
- [42] M. A. Tsfasman, S. G. Vlăduț, and D. Nogin. *Algebraic Geometric Codes: Basic Notions*, volume 139. American Mathematical Soc., 2007.
- [43] P. Vilenkin. On constructions of list-decoding superimposed codes. *Proc. of ACCT-6, Pskov, Russia*, pages 228–231, 1998.
- [44] J. Wolf. Born again group testing: Multiaccess communications. *Information Theory, IEEE Transactions on*, 31(2):185–191, 1985.
- [45] S. Yekhanin. Some new constructions of optimal superimposed designs. In *Proceedings of International Conf. on Algebraic and Combinatorial Coding Theory*, pages 232–235, 1998.
- [46] A. Zhigljavsky. Probabilistic existence theorems in group testing. *Journal of statistical planning and inference*, 115(1):1–43, 2003.